

QC analysis

Jonas Almlöf – Uppsala University

Overview

- Principal component analysis (PCA) of QC variables
- Analysis of individual QC variables
- Expression correlation between samples
- Expression correlation towards QC variables

QC - programs

- FastQC
- RSeQC
- PICARD
- Home made scripts

QC - variables

- Total reads
- Unmapped reads (nr, %)
- Uniquely mapped (nr, %)
- Diff and ratio read-1 and read-2
- Diff and ratio reads map to '+' and '-'
- Non-splice reads (nr, %)
- Splice reads (nr, %)
- Reads mapped in proper pairs (nr, %)
- Clipping_profile bins (nr, %)
- Duplicates (nr, %)
- Optical duplicates (nr, %)
- Estimated library size
- Nr duplicates > 10 (nr, %)
- Gene body coverage bins (nr, %) - Combination of different mRNA sources
- GC in max content (nr, %)
- Total normalized difference of nucleotide content towards normal per cycle
- Total normalized difference of nucleotide content towards normal
- N content per cycle
- %GC
- %GC per cycle
- Mean GC content of sequences Stddev
- Min per base seq quality
- Max per base seq quality
- Max GC percentile
- Mean Q + StdDev
- Fraction N, A, C, G, T + Stddev
- Mean Q for cycles 0-24, 25-49, 50-74 + StdDev
- Fraction bases with Q=2, Q>=10, Q>=20, Q>=30 for cycles 0-24, 25-49, 50-74
- Fraction N, A, C, G, T for cycles 0-24, 25-49, 50-74 + StdDev
- Adapter dimers
- Adapter beginning
- Adapter mid
- Adapter end
- Mean Q30 length + Stddev
- Mean copy number + Stddev
- Presence of 15-mers

QC – outlier analysis

- PCA: Find outliers in the first 4 PC
- Outlier per variable:
Initial filter 3 stdev
Ad hoc filter by importance and effect size

QC – principal components and outliers (BWA)

- PC1 – % GC, % nucleotide content

NA18861.4.M_120208_5, HG00102.3.M_120202_8, NA19247.3.M_120223_7

- PC2 – Quality

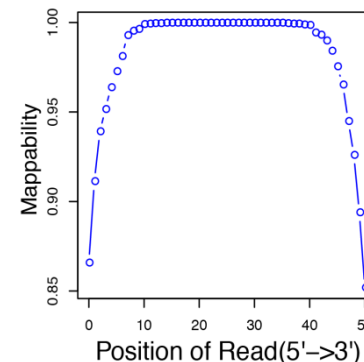
NA12399.7.M_120219_1, HG00139.7.M_120219_1, NA12873.7.M_120219_1, NA20803.7.M_120219_1, NA12889.3.M_120223_7, NA20774.7.M_120219_1, HG00266.6.M_120119_3, HG00263.6.M_120119_3, NA20787.6.M_120119_3

- PC3 – Clipping profile, % gene body coverage

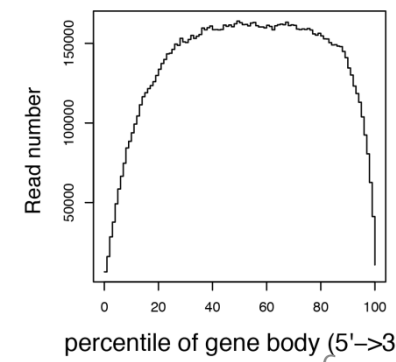
NA20802.1.M_111124_7, NA19153.1.M_111124_7, NA12716.7.M_120219_6, NA18502.7.M_120219_8

- PC4 – N content

NA20773.6.M_120119_4

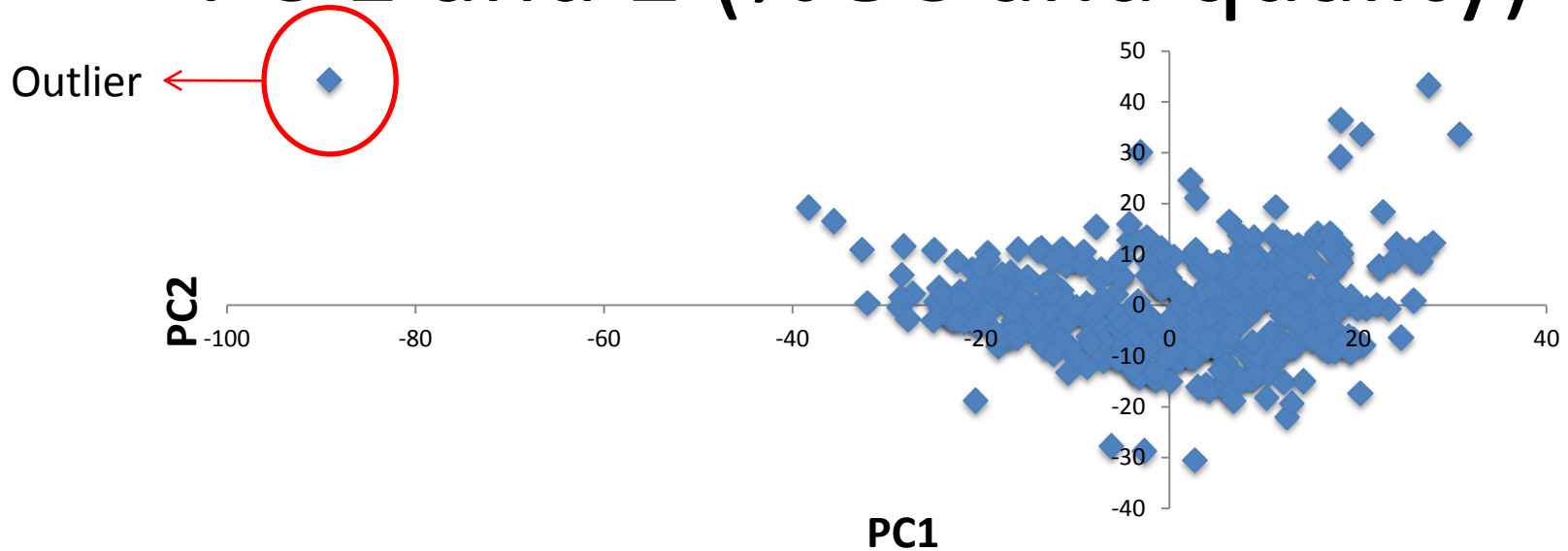


Clipping profile

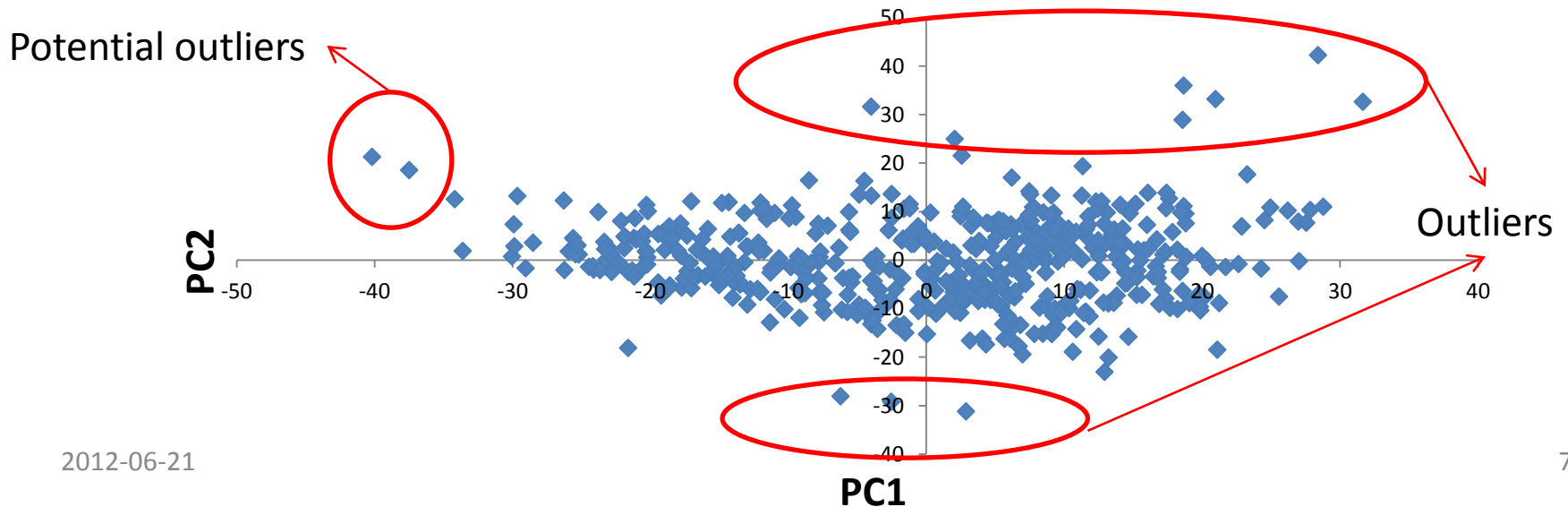


Gene body coverage

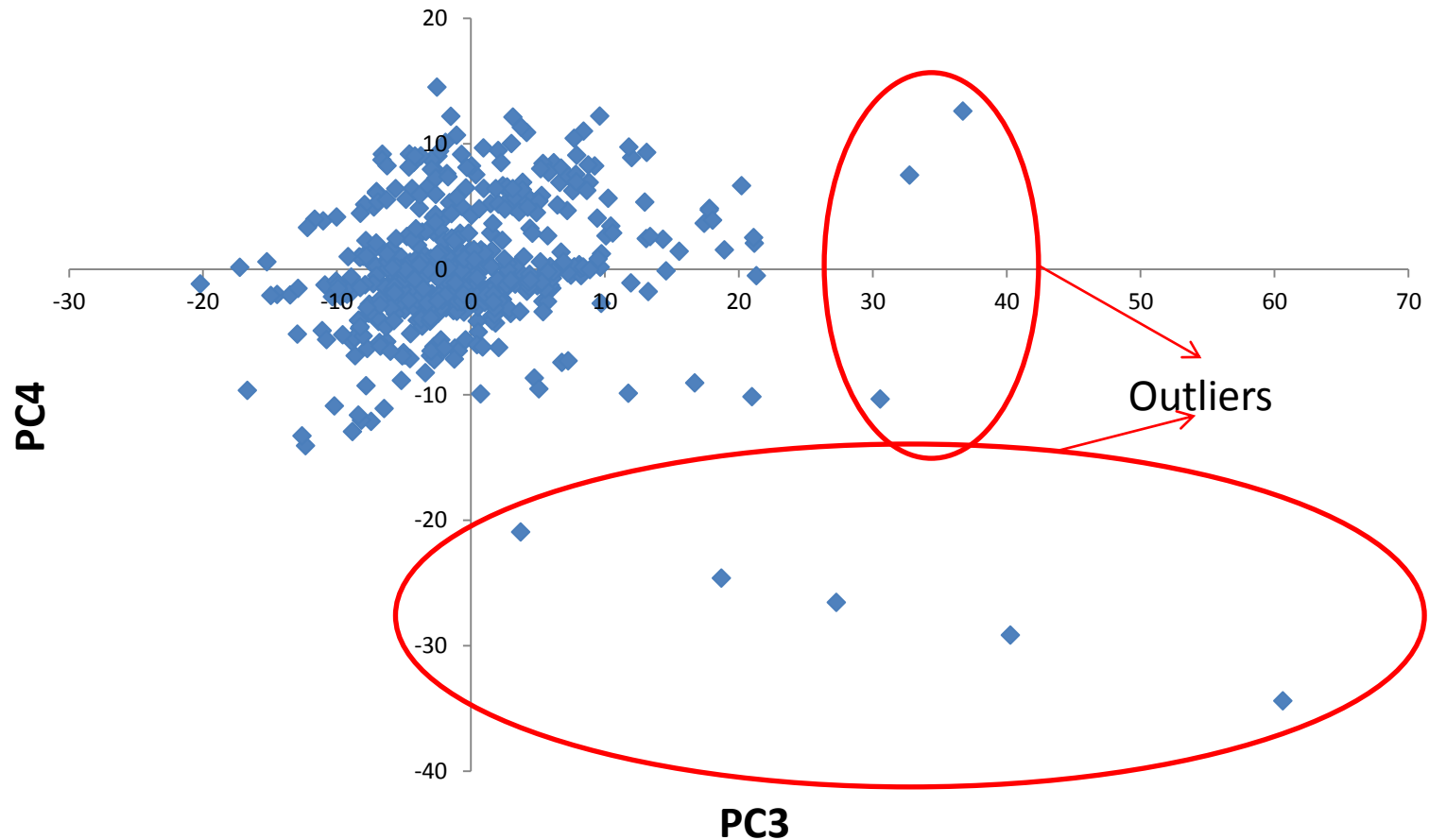
PC 1 and 2 (%GC and quality)



PCA without extreme outlier



PC 3 and 4 (Clipping profile / gene body coverage and N content)

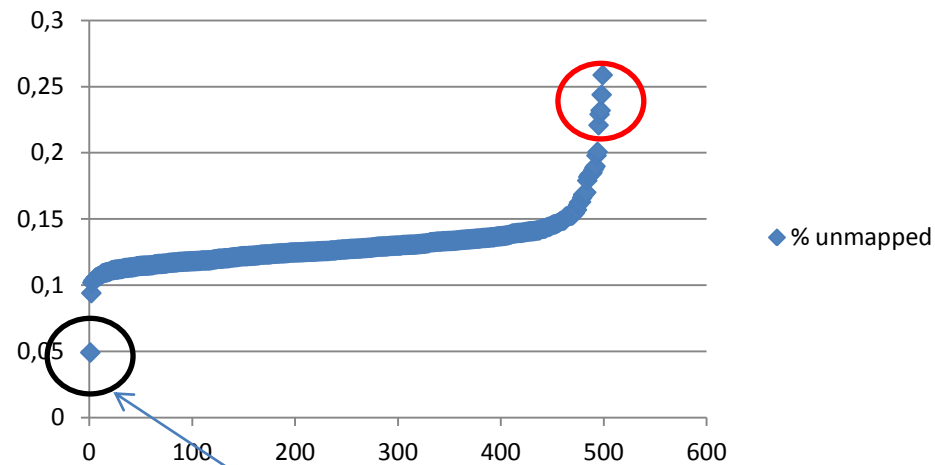


QC – Variable outliers (BWA)

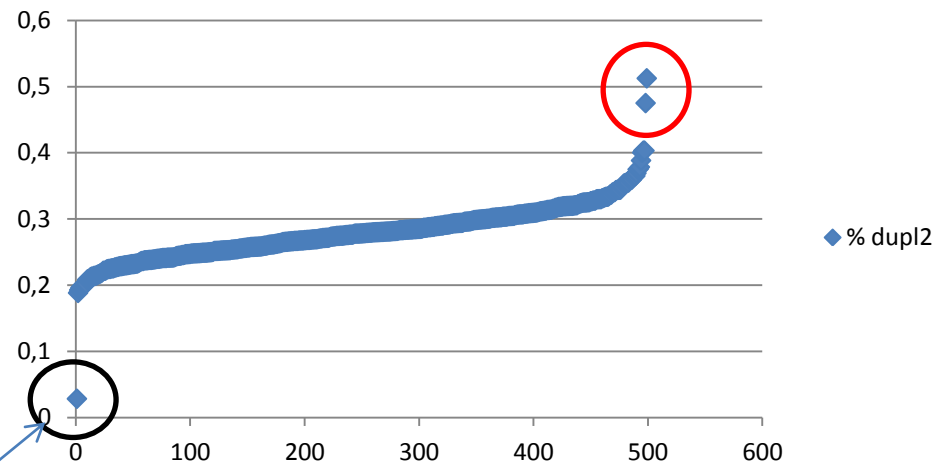
- High read duplication % (Avg: 27.9%, Std: 4.0%)
HG00099.5.M_120131_3 (51%), HG00329.5.M_120131_3 (48%)
- High % unmapped reads (Avg: 13.0%, Std: 1.7%)
NA12889.3.M_120223_7 (26%, PC2), HG00329.5.M_120131_3 (24%, high dupl %),
HG00345.1.M_120209_1 (23%, new), HG00099.5.M_120131_3 (23%, high dupl %),
NA19153.1.M_111124_7 (22%, PC3,4)
- Large difference in number of mapped reads
from read one and two in the paired end read
(Avg: 99.7%, Std: 0.7%)
NA20805.4.M_120208_7 ($r1 / r2 = 0.92$)

QC – Variable outliers cont.

% Unmapped reads

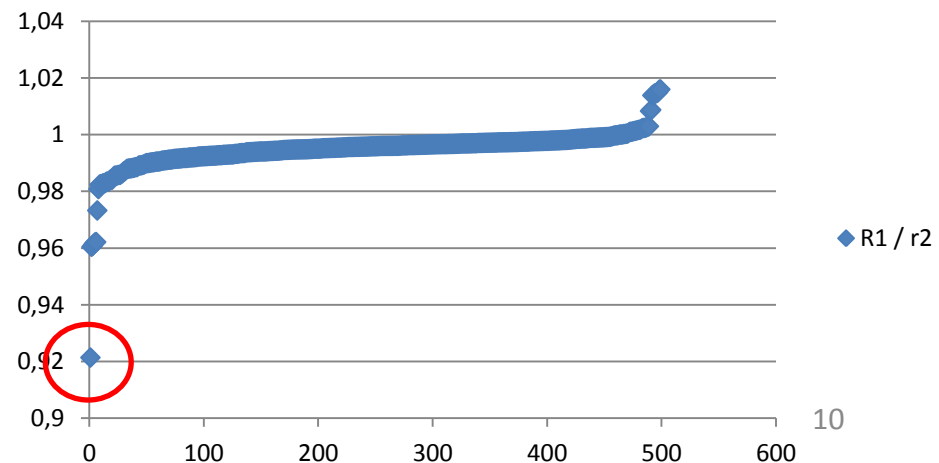


% Duplicates



Are these outliers?

r1 / r2



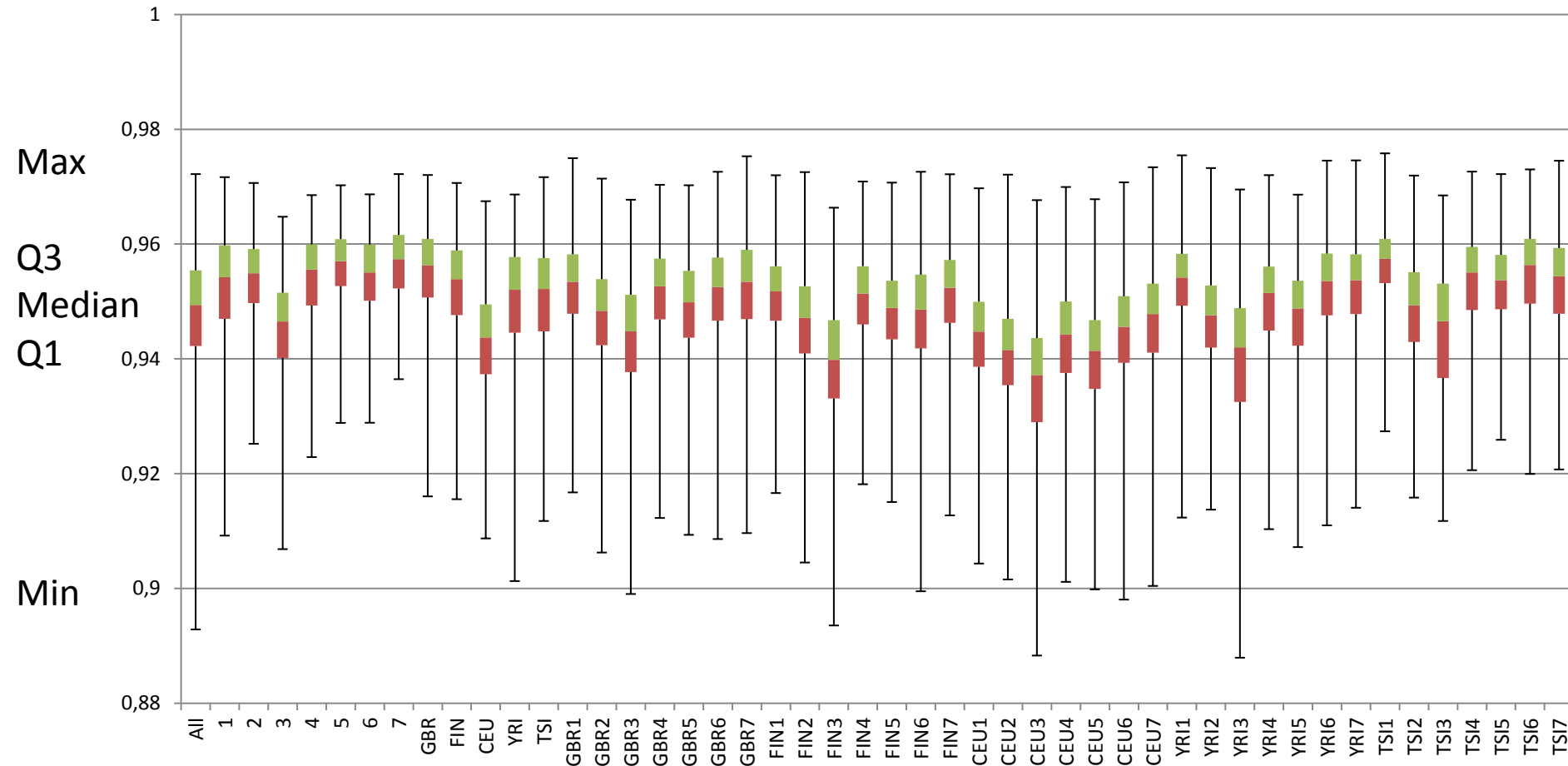
Expression correlation

- Correlate expression in genes and exons between all samples using spearman rank correlation in different setups
- All against all
- Within centers
- Within populations
- Within populations and centers
- Find outliers

Expression correlation – low coverage samples

- Correlations between duplicate samples (low and high coverage of the same sample) to find outliers
- Correlations between duplicate samples should be higher than between different samples

Exon correlation



- Outliers with low average correlation where removed
- Centers and populations differ significantly from All samples (except TSI)

Conclusions

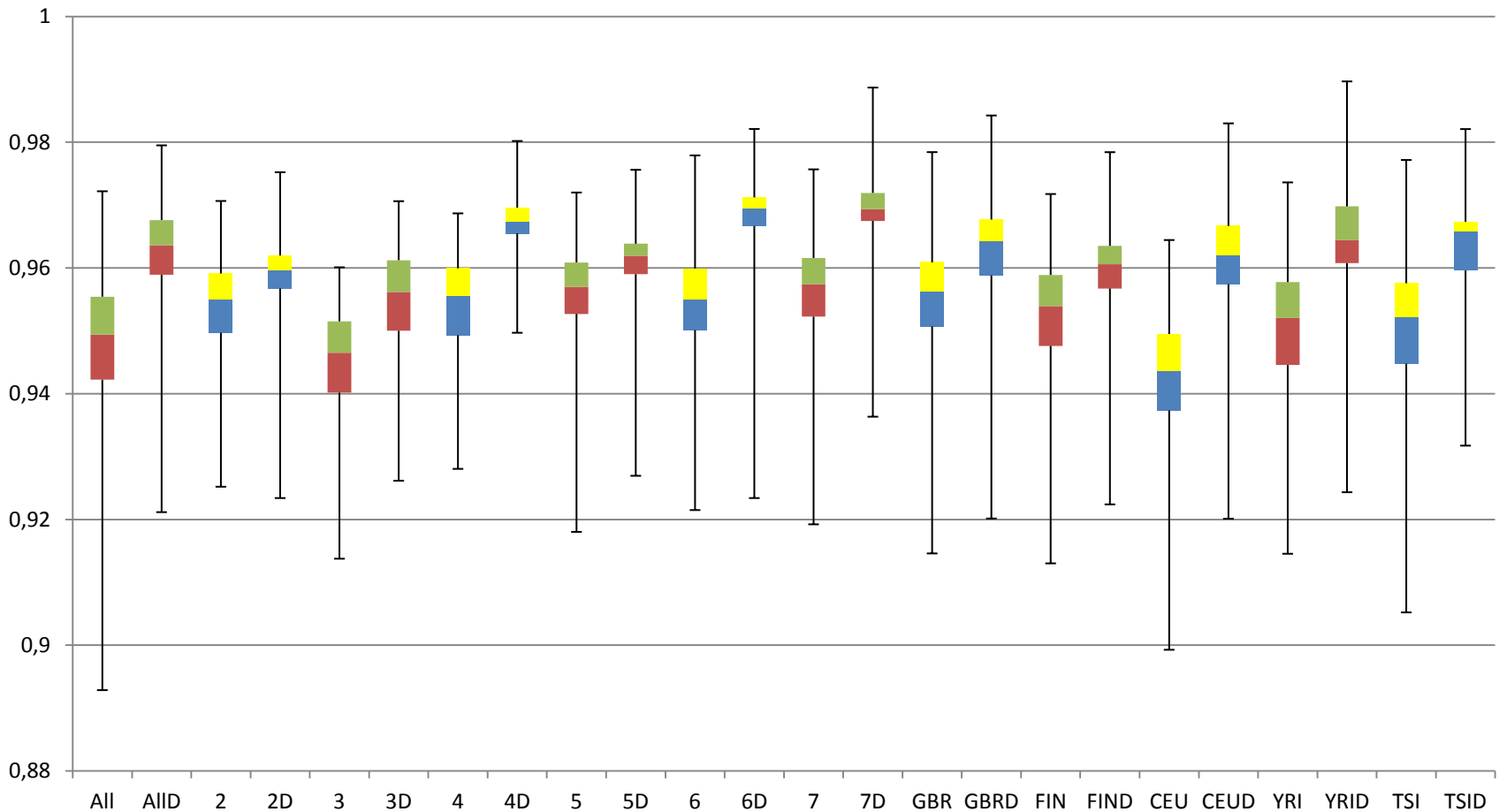
- There are some center effects as within center correlation are significantly higher than for all samples
- Center 3 has lower correlation on average
- Within population have higher correlation than for all samples (except CEU)

Exon correlation outliers / excluded

Sample	Average correlation All: 0.944	Comment
NA19144.4.M_120208_2	0.885	Same as Tuuli found
NA12058.5.M_120131_7	0.920	
NA11931.1.M_111124_8	0.924	
HG00102.3.M_120202_8	0.924	Same as Tuuli found
NA11930.3.M_120202_8	0.926	Same as Tuuli found
NA11892.1.M_111124_2	0.927	
NA19247.3.M_120223_7	0.931	Tuulis outlier (next in line, depends on where you draw the line)
NA19225.6.M_120119_5		Excluded: ASE contamination
NA12399.7.M_120219_1		Excluded: ASE contamination
NA07000.1.M_120209_2		Excluded: ASE contamination
NA18861.4.M_120208_5		Excluded: ASE contamination
HG00237.4.M_120208_1		Excluded: ASE contamination

Exon correlation:

Comparison between different samples and duplicate samples



All within sample correlations are significantly higher

Conclusions

- Correlation between duplicate samples are higher than between different samples => good
Difference large enough? (0.950 vs 0.965)
- The correlation distribution is narrower (Std approximately half) => good
- No center or population outliers

Variable / exon correlation (BWA)

- Clipping profile bins (0.13-0.17 after normalization => 0.08-0.16)
- Estimated library size (0.14 => 0.13)
- % duplicates > 10 (-0.12 => -0.12)
- Gene body coverage 40-50% (0.09-0.13 => 0.10-0.17)
- Gene body coverage 80-100% (-0.11--0.15 => -0.08--0.15)
- % GC in max content (0.15 => 0.13)
- % explained by top 50/100 15 mers (-0.11--0.13 => -0.10--0.12)
- Sum of top correlations (absolute value) 2.65 => 2.29

To do / ideas

- Dubble check QC on GEM-aligned files (analysis is running as we speak)
- Phylogentic trees: compare genotyping data and expression data
- Make more use of the low complexity data